DOCUMENT RESUME

ED 161 940 TM 007 975

AUTHOR Weber, Margaret B.
TITLE The Effect of Choice Format on Internal Consistency.
PUB DATE [77]
NOTE 13p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Academic Achievement; Achievement Tests; Guessing (Tests); *High Achievers; *Low Achievers; *Multiple Choice Tests; *Response Style (Tests); Secondary Education; *Test Construction; Testing Problems; Test Items; *Test Reliability
IDENTIFIERS Distractors (Tests); *Teacher Developed Tests

ABSTRACT

The effects of different choice formats on the reliability of teacher-made tests were examined for high and low achievers. The first study examined the effect of 3 and 5 choice items on the reliability of dichotomously scored teacher-made tests. The second study examined the effect of 3 and 4 choice items on the reliability of similarly designed tests. The results indicated that reliability was not improved by decreasing the number of choices for the high achievers. However, reliability was improved with more choices for low and average achievers. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Margaret B. Weber

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED161940

The Effect of Choice Format on Internal Consistency

Margaret B. Weber

Emory University

TM002 975

Abstract

The purpose of this research was to examine the effect of different choice formats on the reliability of teacher-made tests for high and low achievers. The first study examined the effect of 3 and 5 choice items on the reliability of dichotomously scored teacher-made tests. The second study examined the effect of 3 and 4 choice items on the reliability of similarly designed tests. The results of both studies indicate that reliability was not improved by decreasing the number of choices for the high achievers. However, reliability was improved with more choices for low and average achievers.

The Effect of Choice Format on

Internal Consistency

In a recent discussion, Lord (1977) reviewed four approaches to determining the optimal number of choices per item. The conclusion drawn from the work of Ruch and Stoddard (1925), Ruch and Charles (1928), Williams and Ebel (1957), Tversky (1964) and Grier (1975), is that two or three choices per item result in more reliable tests than do four or five choices given a fixed length of testing time. However, this conclusion may only reflect the increased number of items which can be administered if choices are decreased. In an examination of the relative efficiency of three SAT verbal tests at different ability levels, Lord reported that the effects of the number of choices on reliability given a fixed number of alternatives and a fixed administration time, were not constant across learner differences.

As one would expect, the effect of random guessing becomes more important at low ability levels and less important at high ability levels. Thus, Lord suggests that the reliability issue is specific to learner abilities and should not be assumed constant across differences in ability.

Although Lord's work was done with more precise estimates of low and high ability than are generally available to classroom teachers and with longer tests than those commonly used in instructional settings, the implications of this research are important for test construction at the classroom level. The purpose of this study

was to examine the effect of different choice formats on the reliability of classroom tests across high and low achievers. In order to simulate realistic classroom testing situations, both test length and testing time were held constant.

Two independent studies were conducted to address this issue. The first study examined the effect of 3 and 5 choice items on the reliability of dichotomously scored tests in secondary social science. The second study examined the effect of 3 and 4 choice items on the reliability of dichotomously scored tests in elementary social science.

Study One

Students

Students from two high school geography sections participated in this study. One section was classified as College Preparatory and one section was classified as General Education. These classifications were based on standardized achievement scores earned in eighth grade used in counseling for high school academic programs. The College Preparatory group were ninth graders. The General Education group included 9th through 12th graders. No demographic data was available on these students beyond academic classification.

Instrumentation

One five-choice multiple choice teacher-made test was developed for each of three hierarchically ordered units in high school geography. Each test included knowledge, comprehension application and analysis items (following Bloom, 1956). Content validity[1]

estimates for each test was equal to or greater than .95. A random deletion[2] of two distractors from each item prior to test administration yielded the three choice test.

Procedure

Each student group took the three 25-item tests in both the three-choice and five-choice format. The tests were given in the same order in which the material was taught. The item format order was randomly assigned. Twenty minutes was allowed for completion of each test. This figure, based on teacher judgment, was considered adequate for the majority of both groups to finish all items. A three week interval was maintained between tests I, II, and III. A minimum of three days was maintained between administration of the different forms. Students were told not to guess as the tests were designed for formative evaluation.

Results

As indicated in Table 1, no consistent improvement in internal consistency, as estimated by coefficient α, was achieved by using five-choice items for the low achievers and three-choice items for the high achievers. The reliability estimates in the General

---

Insert Table 1 About Here

Education groups are equal to or higher in the five-choice situation than in the three-choice situation. Thus, this evidence supports the need for more research on the effects of choice format among this learner group.

Study Two

Students

Students from three social science classes participated in this study. These three classes were identified as Average according to overall scholastic achievement of the student groups based on district norms. Strict anonymity was maintained and no additional demographic data was available.

Instrumentation

One four-choice multiple choice test[3] was developed for each student group. Each test included knowledge, comprehension and application items (following Bloom, 1956). Each test had two content validity estimates (see footnote 1) equal to or greater than .95. A random deletion of one distractor from each item prior to administration produced the three-choice test (see footnote 2).

Procedure

The students were systematically, but not randomly, assigned to either test format by seating arrangement. Students were told not to guess as the tests were designed for formative evaluation.

Results

As indicated in Table 2, the coefficient alpha reported for each

---

Insert Table 2 About Here

---

three-choice test was higher than that reported for each corresponding four-choice test. It is important to note that the variance in total test scores was greater in each three-choice test situation

than the corresponding four-choice situation. Given a fixed test length of 19 to 20 items[4], these results suggest that the three-choice item format yields a more efficient or reliable test of average or above average learners than does a four choice item format test. These results are particularly interesting given the limitations on the reliability of short teacher-made tests, the fact that none of these tests had been piloted and given the homogeneous description of the classes.

Conclusion

The results of these two studies suggest that more choices yield higher reliabilities in tests of fixed length and fixed time for low achievers or average achievers. This does not appear to be the case for high achievers. Further research in this area with a wider range of item format differences, subject areas, and learner characteristics is necessary. Clearly, the structure of items and their effect on reliability may be dependent upon selected learner characteristics.

Limitations

The constraints to the generalizability of these results are considerable. The small sample sizes, the length of the tests and their development status are major constraints. Perhaps of a more critical nature, however, is the lack of precision in the ability estimates of each sample. Future research may ameliorate the latter constraint. However, the sample size, test length, and so forth are

characteristic of the application of psychometric findings to teacher-made tests. Recognizing these constraints, it is important to note that classroom teachers have a need for solutions to test construction problems. One vehicle for arriving at these solutions is to conduct psychometric research in classroom settings.

References

1. Bloom, B. S. (ed.) Taxonomy of educational objectives Handbook I: Cognitive domain. New York: David McKay Co., Inc., 1956.
2. Grier, J. B. The number of alternatives for optimum test reliability. Journal of Educational Measurement, 1975, 12, 109-112.
3. Lord, F. M. Optimal number of choices per item--A comparison of four approaches. Journal of Educational Measurement, 14, 1, 1977, 33-38.
4. Lord, F. M. and Novick M. R. Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.
5. Ruch, G. M. and Charles, J. W. A comparison of five types of objective tests in elementary psychology. Journal of Applied Psychology, 1928, 12, 398-404.
6. Ruch, G. M. and Stoddard, G. D. Comparative reliabilities of five types of objective examinations. Journal of Educational Psychology, 1925, 16, 89-103.
7. Tversky, A. On the optimal number of alternatives of a choice point. Journal of Mathematical Psychology, 1964, 1, 386-391.
8. Williams, B. J. and Ebel, R. L. The effect of varying the number of alternatives per item on multiple-choice vocabulary test items. The Fourteenth Yearbook. National Council on Measurements Used in Education, 1957, 63-65.

Footnotes

[1]i.e., percent agreement by competent judges with item classification.

[2]This method of deletion was used as no pilot data was available on the quality of the distractors.

[3]The decision to use four-choice instead of five-choice items in this study follows from Williams and Ebel (1957, 64).

[4]Items with zero variability were excluded from the analysis.

Table 1

College Preparatory Student Group	3 Choice (25 Items)			5 Choice (25 Items)		
	Test I	Test II	Test III	Test I	Test II	Test III
Coefficient Alpha (x)	.81	.62	.74	.80	.67	.84
$\bar{X}$	18.20	14.43	14.80	15.92	15.38	16.45
σ^2	19.43	12.99	14.38	20.58	14.26	25.67
N	20	28	20	20	28	20
General Education Student Group						
Coefficient Alpha (x)	.76	.74	.75	..81	.67	.90
$\bar{X}$	17.33	16.87	14.81	16.38	15.73	16.73
σ^2	17.54	12.57	18.70	17.76	13.07	35.42
N	13	13	11	13	13	11

Table 2

Average Students	3-Choice	4-Choice
Class I (19 items)	α = .69	α = .50
	$\bar{X}$ = 11.00	$\bar{X}$ = 13.40
	s^2 = 13.21	σ^2 = 7.16
	N = 17	N = 19
Class II (20 items)	α = .78	α = .57
	$\bar{X}$ = 13.75	$\bar{X}$ = 14.73
	S^2 = 13.68	S^2 = 6.68
	N = 28	N = 20
Class III (19 items)	α = .59	α = .10
	$\bar{X}$ = 11.08	$\bar{X}$ = 9.08
	S^2 = 10.58	S^2 = 4.91
	N = 13	N = 13